frontiers in Microbiology

MINI REVIEW
published: 06 August 2019
doi: 10.3389/fmicb.2019.01722

Machine Learning Approaches for Epidemiological Investigations of Food-Borne Disease Outbreaks

Baiba Vilne[1,2]*, *Irēna Meistere*[1], *Lelde Grantiņa-Ieviņa*[1] and *Juris Ķibilds*[1]

[1] Institute of Food Safety, Animal Health and Environment—"BIOR," Riga, Latvia, [2] SIA net-OMICS, Riga, Latvia

Foodborne diseases (FBDs) are infections of the gastrointestinal tract caused by foodborne pathogens (FBPs) such as bacteria [*Salmonella, Listeria monocytogenes* and Shiga toxin-producing *E. coli* (STEC)] and several viruses, but also parasites and some fungi. Artificial intelligence (AI) and its sub-discipline machine learning (ML) are re-emerging and gaining an ever increasing popularity in the scientific community and industry, and could lead to actionable knowledge in diverse ranges of sectors including epidemiological investigations of FBD outbreaks and antimicrobial resistance (AMR). As genotyping using whole-genome sequencing (WGS) is becoming more accessible and affordable, it is increasingly used as a routine tool for the detection of pathogens, and has the potential to differentiate between outbreak strains that are closely related, identify virulence/resistance genes and provide improved understanding of transmission events within hours to days. In most cases, the computational pipeline of WGS data analysis can be divided into four (though, not necessarily consecutive) major steps: *de novo* genome assembly, genome characterization, comparative genomics, and inference of phylogeny or phylogenomics. In each step, ML could be used to increase the speed and potentially the accuracy (provided increasing amounts of high-quality input data) of identification of the source of ongoing outbreaks, leading to more efficient treatment and prevention of additional cases. In this review, we explore whether ML or any other form of AI algorithms have already been proposed for the respective tasks and compare those with mechanistic model-based approaches.

Keywords: machine learning, food-borne disease, outbreaks, bacterial WGS, bioinformatics analysis pipeline

OPEN ACCESS

Edited by:
Sophia Johler,
University of Zurich, Switzerland

Reviewed by:
Laura M. Carroll,
Cornell University, United States
Heather A. Carleton,
Centers for Disease Control and Prevention (CDC), United States

***Correspondence:**
Baiba Vilne
baiba.vilne@bior.lv

Specialty section:
This article was submitted to Food Microbiology, a section of the journal Frontiers in Microbiology

Received: *07 March 2019*
Accepted: *12 July 2019*
Published: *06 August 2019*

Citation:
Vilne B, Meistere I, Grantiņa-Ieviņa L and Ķibilds J (2019) Machine Learning Approaches for Epidemiological Investigations of Food-Borne Disease Outbreaks. Front. Microbiol. 10:1722. doi: 10.3389/fmicb.2019.01722

1. INTRODUCTION

Foodborne diseases (FBDs) are infections of the gastrointestinal tract caused by foodborne pathogens (FBPs) such as bacteria and several viruses, but also parasites and some fungi. *Salmonella, Listeria monocytogenes* and Shiga toxin-producing *Escherichia coli* (STEC) are some of the most important bacterial FBPs (Sekse et al., 2017), causing the most outbreaks and the largest number of sporadic cases with severe illness or even fatal outcome (EFSA, 2015; Sekse et al., 2017). *Salmonella* infections affect people at all ages and the main food sources of infection typically include ready-to-eat foods, eggs, swine and poultry. *L. monocytogenes* infections mostly affect elderly people, as well as immunocompromised patients and pregnant women, and display high mortality rates. Common food sources of *L. monocytogenes* include ready-to-eat foods such as smoked fish and soft cheeses. STEC has been associated with severe complications, e.g., acute kidney failure, often affecting elderly and immunocompromised people, and also small children.

The main food sources of STEC infections are bovine meat, followed by vegetables and juice (EFSA, 2015).

Whole-genome sequencing (WGS) is becoming more accessible and affordable as a routine approach for early detection of FBD outbreaks (Buultjens et al., 2017; Sekse et al., 2017). WGS captures the entire genome within hours to days and has the potential to differentiate between outbreak strains that are closely related, identify virulence/resistance genes and provide improved understanding of transmission events (Quainoo et al., 2017; Andersen and Hoorfar, 2018). Moreover, third-generation sequencing technologies such as Oxford Nanopore (ONT) sequencing and PacBio Single Molecule, Real-Time (SMRT), which allow the generation of ultra-long (up to 300 kb) reads, are well suited to assemble reference genomes from outbreak strains *de novo*, potentially contributing to more precise taxonomic assignment, while offering increased detection speed and relatively decreasing costs, as, in comparison to Illumina short-read sequencing, both technologies are still three and almost seven times more expensive, respectively (Brown et al., 2017; Sekse et al., 2017; Nicola De Maio, 2019). Several *proof-of-concept* studies have demonstrated the superiority of WGS over traditional typing methods for a range of high priority food-borne pathogens, e.g., *Salmonella enterica, Listeria monocytogenes, Campylobacter species* and STEC (Kanamori et al., 2015; Quick et al., 2015; Moran-Gilad, 2017). Large initiatives have emerged to investigate the options of replacing conventional methods with WGS for outbreak investigations. Two such examples include the ENGAGE (Establishing Next Generation sequencing Ability for Genomic analysis in Europe) (Hendriksen et al., 2018) and INNUENDO projects (Llarena et al., 2018), focusing on the idevelopment of dedicated analytical platforms and standardized analysis pipelines, e.g., for *E. coli* and different *Salmonella* spp. serotypes (Hendriksen et al., 2018).

In the era of Big Data, as the volume and complexity of data increases steadily, artificial intelligence (AI) and its sub-discipline machine learning (ML) are re-emerging and gaining an ever increasing popularity in the scientific community and industry (Ching et al., 2018). While mechanistic model-based approaches aim at constructing simplified mathematical formulations, i.e., hypothesis, of causal mechanisms by carefully observating, analyzing and trying to understand the complexity of the respective phenomenon (Baker et al., 2018), machine learning (ML) algorithms use large-scale datasets to extract meaningful patterns (i.e., "learn") and use this "knowledge" to make predictions on other data (Alkema et al., 2016). Moreover, ML can be done in a unsupervised manner by exploring and detecting patterns within the data or in a supervised manner by classifying, predicting and explaining (Tebani et al., 2016). Unsupervised ML techniques involve well-known and widely used methods such as principal component analysis (PCA) and k-means clustering (Tebani et al., 2016). PCA is a dimensionality reduction method, transforming a large set of variables into a smaller set, while preserving as much information as possible (Hotelling, 1933), whereas k-means clustering groups similar data points together in a fixed number (k) of clusters and tries to discover their underlying patterns (Hartigan and Wong, 1979). In life sciences, some frequently used supervised ML strategies have been Random Forest (RF), Support Vector Machines (SVM), Naive Bayes (NB), and Artificial Neural Networks (Lai et al., 2016). RF alorithm randomly selects a subset from the training data to construct an ensemble of decision tree predictors to aggregate the predictions, thus lowering the variance (Breiman, 1996). SVM represent a pattern classification technique, which is based on the idea of transforming the original data that is not linearly separable to a higher dimensional space and finding a hyperplane separating the data into classes (Boser et al., 1992). NB represents a probabilistic algorithm that uses the probability theory and Bayes' Theorem in conjunction with prior knowledge to calculate the probability of each feature to belong to each of the classes and then outputs the class with the highest probability (Devroye et al., 2013). Finally, ANNs are graph computing models, which, at least to some extent, should mimic the functioning of the human brain, hence its computing units are called neurons and are interconnected for passing information to each other. Moreover, networks of neurons are additionally organized in layers. The first one is an input layer, receiving the training data. This is followed by several hidden layers. The last one is an output layer, which performs the actual prediction of the class (Kruse et al., 2016).

Global multi-disciplinary initiatives like One Health (OH) (http://www.onehealthinitiative.com/), aiming toward optimizing the health of people, animals and the environment, would greatly profit from such approaches, as multiple complex challenges need to be addressed, including the maintenance of a safe food and water supply for a growing human population. Considering the current ease with which people and animals or animal products can be transported around the globe, the forefront issues of OH are clearly related to spread of emerging infectious diseases and antimicrobial resistance (AMR) (Gibbs, 2014). Especially, outbreaks caused by multi-drug-resistant bacteria are an urgent and growing global public health threat (CDC, 2013; WHO, 2014). Effective management protocols must be in place, as quick identification leads to faster and more precisely targeted treatment (Quainoo et al., 2017).

ML strategies have already been used for microbial diagnostics in diverse contexts, including (i) taxonomic grouping of metagenomics data (Sedlar et al., 2017; Afify and Al-Masni, 2018); (ii) classification of *L. monocytogenes* persistence in retail delicatessen environments (Vangay et al., 2014); (iii) phenotype prediction of bacterial strains based on presence/absence of particular genes (i.e., gene-trait matching) (Dutilh et al., 2013; Alkema et al., 2016; Farrell et al., 2018); (iv) to identify strains that demonstrate a higher probability to cause severe diseases (Wheeler et al., 2018); (v) to predict the host range of pathogens (Lupolova et al., 2017), e.g., identifying their signatures of host adaptation (Wheeler et al., 2018); and (vi) to predict the antimicrobial resistance potential of different *E. coli* strains (Her and Wu, 2018) or from different sources (Li et al., 2018).

The WGS data analysis pipeline can be generally divided into four major steps (**Figure 1**): *de novo* genome assembly, genome characterization, comparative genomics and inference of phylogeny or phylogenomics (Quainoo et al., 2017). However, these steps are not necessarily consecutive, depending on the

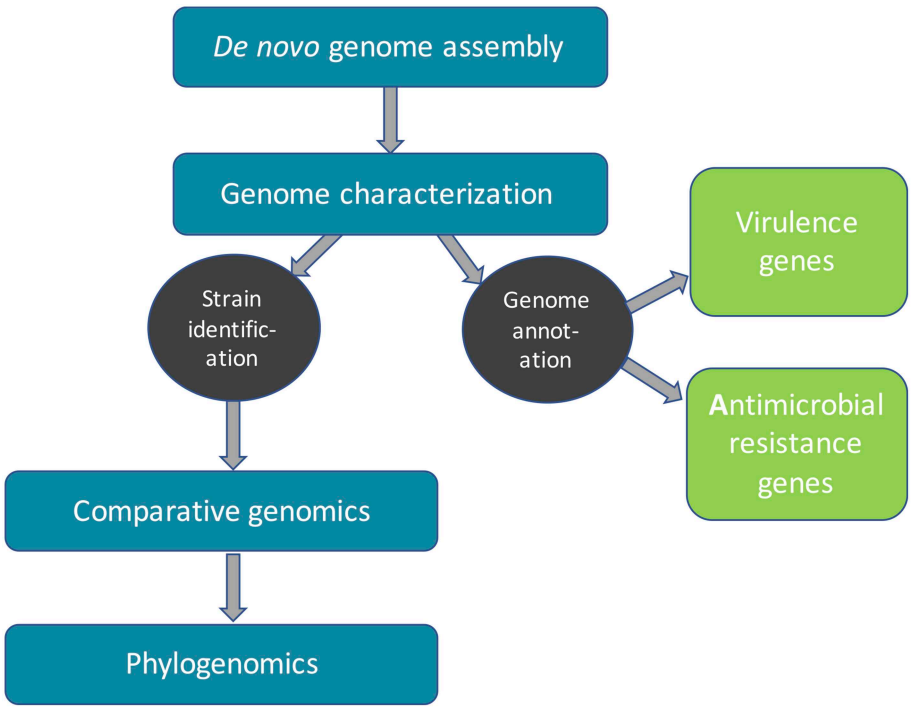

FIGURE 1 | An overview of an example bacterial sequence data analysis workflow.

objectives of the study. ML could be used in any of these analyses to increase the speed and potentially accuracy (provided increasing amounts of high-quality input data). In this review, we aim to explore whether ML algorithms have already been proposed for the respective task and compare those algorithms with mechanistic model-based approaches (see **Table 1** for an overview). We mainly focus on single-genome short-read (Illumina) bacterial WGS; however in cases where, to the best of our knowledge, no ML algorithms have been reported for the respective task, we also briefly touch upon ML algorithms dedicated to ultra-long read technologies, 16S metataxonomics and shotgun metagenomics, as these approaches may find future applications in FBD outbreaks. Currently, the starting point for any FBD outbreak investigation involving strain typing is access to isolates, which may be difficult to obtain or are often even unavailable. Moreover, most food samples are complex, harboring composite microbial communities. In this regard, metagenomic approaches would allow one to capture the full spectrum of microbes in foods entirely without prior need for culturing and isolation, allowing also the detection of "viable but not cultivable," as well as non-viable microbes (Bergholz et al., 2014).

2. MACHINE LEARNING FOR *DE NOVO* MICROBIAL GENOME ASSEMBLY

Genome assembly tools are applied with the purpose of assembling the sequencing reads into larger fragments (i.e., contigs), from which near-complete genomes can be further re-constructed. As the read lengths of the second generation (e.g., Illumina) technologies are short (i.e., 50–300 bp), *de novo* assembly without a reference genome remains a challenging task (Zhu et al., 2014). However, *de novo* assembly is especially relevant in FBD outbreak investigations, where the source strain might be undetectable with conventional methods and thus taxonomically unclassified (Quainoo et al., 2017). Currently, the majority of the algorithms are based on the de Bruijn graph or overlap-layout strategies. The de Bruijn graph algorithm first splits up each read into smaller substrings, k-mers, which are further used to construct a graph, in which k-mers represent nodes; two nodes are connected with an edge if they overlap by k-1 nucleotides and follow each other in the read. Thus, each contig is represented as a path within the graph (Zhu et al., 2014). The overlap-layout-based algorithms start by computing the overlaps among all the reads, which are then used to perform the genome assembly (Zhu et al., 2014). For short Illumina read-based single genome WGS, the most popular assemblers include Velvet (Zerbino and Birney, 2008), IDBA-UD (Peng et al., 2012), RAY (Boisvert et al., 2010), SPAdes (Bankevich et al., 2012), and SKESA (Souvorov et al., 2018), all of which employ the de Bruijn graph-based assembly strategy. The overlap-layout-based algorithms are mainly used for the assembly of ultra-long reads: Minimap/miniasm (Li, 2016) and Canu (Koren et al., 2017).

For 16S metataxonomics data, interestingly, there is a tool REAGO (REconstruct 16S ribosomal RNA Genes from

TABLE 1 | An non-exhaustive list of the mechanistic model-based vs. ML tools for microbial genome analysis.

Category	Tools	
	Mechanistic model-based	**Machine learning**
***DE NOVO* GENOME ASSEMBLY**		
	Velvet (Zerbino and Birney, 2008), IDBA-UD (Peng et al., 2012), RAY (Boisvert et al., 2010), SPAdes (Bankevich et al., 2012), SKESA (Souvorov et al., 2018) Minimap/miniasm (Li, 2016), Canu (Koren et al., 2017), REAGO** (Yuan et al., 2015)	PERGA (Zhu et al., 2014), Minimus/AMOS (Palmer et al., 2010), MetaVelvet-SL* (Cheng, 2015)
GENOME CHARACTERIZATION		
1. Bacterial strain identification	BLASTN (McGinnis and Madden, 2004), JSpeciesWS (Richter et al., 2016), ANIttools (Han et al., 2016), OrthoANI (Lee et al., 2016), KmerFinder (Hasman et al., 2014), StrainSeeker (Roosaare et al., 2017), MESH (Ondov et al., 2016), Kraken* (Wood and Salzberg, 2014), MetaPhlAn* (Segata et al., 2012), QIIME2** (Caporaso et al., 2010), MOTHUR** (Schloss et al., 2009), MG-RAST** (Meyer et al., 2008)	PaPrBaG (Deneke et al., 2017), NBC (Rosen et al., 2008), TACOA (Diaz et al., 2009), PhyloPythiaS+* (McHardy et al., 2007; Gregor et al., 2016), BLCA** (Gao et al., 2017), 16S Classifier** (Chaudhary et al., 2015)
2. Bacterial genome annotation	PROKKA (Seemann, 2014), RAST/myRAST (Overbeek et al., 2014), MetaGeneAnnotator* (Noguchi et al., 2008), MetaGene* (Noguchi et al., 2006), Tax4Fun** (Aßhauer et al., 2015)	Woods (Sharma et al., 2015), Orphelia* (Hoff et al., 2009), MGC* (El Allali and Rose, 2013), MetaGUN* (Liu et al., 2013), Meta-MFDL* (Chen et al., 2016)
3. Virulence gene detection	VirulenceFinder (Joensen et al., 2014), PathogenFinder (Cosentino et al., 2013)	BacFier (Iraola et al., 2012), PaPrBaG (Deneke et al., 2017)
4. Antimicrobial resistance gene detection	ResFinder (Zankari et al., 2012), RGI/CARD (Jia et al., 2017), AMRFinder (Feldgarden et al., 2019)	DeepARG (Arango-Argoty et al., 2018), PATRIC (Antonopoulos et al., 2017)
COMPARATIVE GENOMICS		
1. Reference-based SNP methods	CSI Phylogeny (Kaas et al., 2014), Lyve-SET (Katz et al., 2017), CFSAN SNP Pipeline (Davis et al., 2015), SPANDx (Sarovich and Price, 2014), SNVPhyl (Petkau et al., 2017)	
2. Non-reference-based SNP analysis	KSNP (Gardner et al., 2015)	
3. Pangenome-based analysis	Roary (Page et al., 2015), PanWeb (Pantoja et al., 2017), Pan-Seq (Laing et al., 2010)	
4. Core genome/whole-genome multi-locus sequence typing (MLST)	EnteroBase (Alikhan et al., 2018), BIGSdb (Jolley and Maiden, 2010), chewBBACA (Silva et al., 2018)	BAPS/hierBAPS (Cheng et al., 2011, 2013)
PHYLOGENOMICS		
	RAxML (Stamatakis et al., 2005), FastTree (Price et al., 2009), CSI Phylogeny (Kanamori et al., 2015), Lyve-SET (Katz et al., 2017), PHYLIP (Shimada and Nishida, 2017), BEAST (Drummond and Rambaut, 2007)	

**The tool is dedicated to shotgun metagenomics; ** the tool dedicated to 16S metataxonomics.*

metagenOmic data), which combines homology search that considers also the secondary structure and properties of 16S ribosomal RNA genes to perform their *de novo* reconstruction (Yuan et al., 2015).

ML has been used in PERGA (Paired-End Reads Guided Assembler) (Zhu et al., 2014) to determine the correct contig extension. For this, the alogirthm constructs a decision model, considers the avaialble information from paired-end reads such as different read overlap size and various branch features, i.e., path weight, read coverage levels and gap size. In addition, PERGA also detects tandem repeats with the aim to resolve branches in the assembly graph and construct longer and more accurate contigs and scaffolds (Zhu et al., 2014). Minimus/AMOS (Palmer et al., 2010) contains a module that uses ML (C4.5 decision tree, NB and RF) in combination with features identified from prior sequencing projects and completed genomes to classify overlaps as true or false, by this improving the quality of the genome assembly.

For shotgun metagenomics, ML-based strategies has been proposed in order to pre-allocate (i.e., cluster) reads into similar groups before the assembly step, thus reducing the overall computational complexity of the process (Cheng, 2015). Moreover, when assembling metagenomics data, the de Bruijn graph is usually decomposed into individual sub-graphs to build an isolated genome; however, there are still the so called chimeric nodes, i.e., those present in more than one sub-graph, which need to be identified and split apart (Afiahayati et al., 2015). For this, ML (SVM) has been applied, e.g., as implemented in MetaVelvet-SL (Afiahayati et al., 2015).

3. MACHINE LEARNING FOR MICROBIAL GENOME CHARACTERIZATION

After assembly, the bacterial identity of the isolate usually needs to be identified, followed by genome annotation and identification of those genes that might be of clinical importance,

such as antimicrobial resistance and virulence genes. For this, genome characterization tools are being developed which compare the assembled contigs to several reference databases of known genes and reference genomes (Quainoo et al., 2017).

3.1. Bacterial Strain Identification

In this category, computational tools, which can assess bacterial identity either directly from reads or from pre-assebled contigs are used (Quainoo et al., 2017). Current tools are often based on genome-wide sequence similarity statistics (Ciufo et al., 2018). NCBI BLAST (the Basic Local Alignment Search Tool) is one of the most popular alignment tools and its variant BLASTN can be used to identify species from contigs using the Nucleotide Collection (nr/nt) database, which contains all the microbial sequences from the NCBI database (McGinnis and Madden, 2004). However, for large-scale read mapping, BLAST may be too slow (Deneke et al., 2017). Generally, this approach may fail to detect novel species in cases when closely related genomes are not found in the reference databases (Deneke et al., 2017), which are known to be biased toward cultivable pathogenic bacteria (Farrell et al., 2018). Average Nucleotide Identity (ANI) (Clingenpeel et al., 2015) has been recently proposed as an alternative metrics for the identification and classification of bacterial species, calculated by performing several pair-wise comparisons of all sequences shared between two given strains. This method is implemented within tools such as JSpeciesWS (Richter et al., 2016), ANItools (Han et al., 2016), and OrthoANI (Lee et al., 2016). Alternatively, composition-based methods such as KmerFinder (Hasman et al., 2014) exist, which employ a precomputed database compiled using 1,647 complete bacterial genomes from the NCBI database divided into 16-mers. Given an input file of unknown bacterial species, the program provides an overview of all k-mers that match all the templates in the database (i.e., the "standard" method) or counts all the k-mers that might originate from a particular strain (i.e., the "winner takes it all" method; Hasman et al., 2014). StrainSeeker (Roosaare et al., 2017) starts with a Newick-format tree and derives a list of k-mers for each node in that tree. Thereafter, the observed vs. expected fractions of node-specific k-mers are being analyzed to determine each node's presence in the input data (Roosaare et al., 2017). MESH (Ondov et al., 2016) is another k-mer based strain identification algorithm that extends the MinHash dimensionality-reduction technique by reducing large (sets of) sequences into small, representative sketches, which are then used to infer global mutation distances.

For shotgun metagenomics, Kraken (Wood and Salzberg, 2014) is a k-mer based approach, which tries to match 31-mers from the input data to a pre-computed database, by considering all reference genomes in which they occur and then mapping these 31-mers to the lowest common ancestor. MetaPhlAn (Segata et al., 2012) first collects all clade-specific marker genes, i.e., from strain to phylum, into a database, which it then utilized for the taxonomic classification of metagenomic shotgun data.

For 16S metataxonomics data, sequence alignment-based approaches are usually used to assign taxa (Chaudhary et al., 2015). For this, QIIME2 (Caporaso et al., 2010), MOTHUR (Schloss et al., 2009), and MG-RAST (Meyer et al., 2008) are the most commonly used pipelines. Overall, the major limitations of the above approaches are the computational time requirements and dependence on the reference databases (Chaudhary et al., 2015).

To overcome these limitations, ML-based approaches have been proposed. NBC (Rosen et al., 2008) calculates k-mer frequency profiles of all publicly available microbial reference genomes and uses these profiles to train a naive Bayesian classifier to identify the respective genome by any query fragment. TACOA (Diaz et al., 2009) achieves taxonomic classification by combining the k-nearest neighbor algorithms with kernel-based ML strategies. Yet another ML-based approach, PaPrBaG (Pathogenicity Prediction for Bacterial Genomes), has been recently proposed, which, in addition to taxonomic classification, also aims to predict the pathogenic potential of the respective strains (Deneke et al., 2017).

For shotgun metagenomics, PhyloPythiaS+ (McHardy et al., 2007; Gregor et al., 2016) is a sequence composition-based method that uses hierarchical structured-output by employing a multiclass support vector machine (SVM) classifier.

For 16S metataxonomics data, prediction-based ML approaches for taxonomic classification have started to emerge, as opposed to homology-based methods (Chaudhary et al., 2015). For example, BLCA is a tool for taxonomic classification of 16S rRNA gene sequences, which combines sequence similarity to the reference database with Bayesian posterior probabilities to weight the degree of sequence similarity of the query sequence to every hit from the database (Gao et al., 2017). 16S Classifier is a similar tool that deploys RF and is compatible with the QIIME2 pipeline (Chaudhary et al., 2015).

3.2. Bacterial Genome Annotation

Bacterial genome annotation tools explore which genes are contained in the respective bacterial genome by retrieving the relevant features (i.e., coding regions and their putative products, non-coding RNAs and signal peptides) from raw reads or pre-assembled contigs (Seemann, 2014; Quainoo et al., 2017). PROKKA (Seemann, 2014) is a software suite unifying several feature prediction tools, such as Prodigal (Hyatt et al., 2010) for the identification of coding sequences, RNAmmer (Lagesen et al., 2007), Aragorn (Laslett and Canback, 2004), and Infernal (Kolbe and Eddy, 2011) for the prediction of ribosomal, transfer and non-coding RNA genes, respectively, as well as SignalP (Petersen et al., 2011) to identify signal leader peptides. RAST/myRAST (Overbeek et al., 2014) is another popular genome annotation tool, which uses a SEED k-mer-based annotation algorithm to predict coding sequences, as well as tRNAs and rRNAs.

For shotgun metagenomics, there are several model-based approaches, including MetaGeneAnnotator (Noguchi et al., 2008) or MetaGene (Noguchi et al., 2006), both using Markov chain models to identify genes.

However, the main limitation of these models is that they require optimization of thousands of parameters, which limits their practical use (Zhang et al., 2017). Sequence similarity-based methods, on the other hand, are considered rather time-consuming and computationally demanding, especially when applied to shotgun metagenomic data. This poses a bottleneck

for efficient sequencing data analysis (Sharma et al., 2015). Moreover, RAST is known to have difficulties dealing with mixed or contaminated cultures, as its algorithm relies on closely related isolates (Quainoo et al., 2017). In addition, these methods are used to find genes with previously known homologous proteins and cannot predict novel genes (Zhang et al., 2017).

Unfortunately, 16S metataxonomic data does not provide any information on functional genes and proteins for the microbial communities being analyzed (Aßhauer et al., 2015); however, these can be predicted using pangenome-based approaches such as Tax4Fun (Aßhauer et al., 2015).

Alternatively, ML (RF) and similarity-based (RAPsearch2) approaches have been combined in a tool called "Woods" (Sharma et al., 2015); however, it is currently restricted to the prediction of protein coding sequences only.

For shotgun metagenomics, several ML-based methods have been proposed, such as Orphelia (Hoff et al., 2009), MGC (El Allali and Rose, 2013), MetaGUN (Liu et al., 2013), and Meta-MFDL (Zhang et al., 2017), e.g., the latter using a deep stacking networks learning model and multiple genomic features (i.e., the usage of monocodons and monoamino acids) for identifying genes from metagenomic fragments (Zhang et al., 2017).

3.3. Virulence Gene Detection

In this part of the analysis, the aim is to explore whether the previously annotated genes infer virulence, i.e., some degree of pathogenicity to the host (Quainoo et al., 2017). However, virulence gene detection does not necessarily have to follow the genome annotation step. It can also be performed either using reference database entries as BLAST queries against assembled genomes or mapping raw reads against reference database entries (or any other collection of genes of interest). Also, predicted (but not annotated) coding DNA (or predicted protein) sequences can be screened for virulence gene content. The most commonly used reference database for virulence genes is the Virulence Factor Database (VFDB) (Chen et al., 2016), containing information on 951 bacterial strains and 1,075 virulence factors (as of March 2019), including different characteristics, such as whether a virulence factor is used in offensive or defensive actions. Recently, VFDB has been supplemented with VFanalyzer, a Web-based tool that builds orthologous groups of genes using a query genome and pre-analyzed reference genomes and then performs sequence similarity searches among the VFDB gene collection for atypical and strain-specific virulence genes (https://doi.org/10.1093/nar/gky1080). Frequently used tools to predict virulence genes from sequencing data include VirulenceFinder (Joensen et al., 2014), a Web-based tool that uses BLASTN (Camacho et al., 2009) and contains virulence markers for four microbes: *Listeria, S. aureus, E. coli*, and *Enterococcus*. Another Web-based tool is PathogenFinder (Cosentino et al., 2013), which assumes that bacterial pathogenicity (or lack of it) depends on groups of proteins that are consistently found together in either pathogens or non-pathogens. PathogenFinder aims to identify such groups of proteins.

Several ML-based approaches have been proposed for virulence gene detection. VirulentPred (Garg and Gupta, 2008) is a bi-layer cascade SVM-based prediction method, where the first layer classifiers are being trained using different protein sequence features, such as amino acid and dipeptide composition. The results from the first layer are then passed to the second layer classifier, which utilizes sequence similarity and a BLAST database containing both virulence and non-virulence genes. BacFier (Iraola et al., 2012) uses known pathogenic vs. non-pathogenic strains and their genetic features (e.g., the presence or absence of different virulence-related genes) to train ML algorithms in predicting pathogenicity of input bacterial genomes. Finally, as described above, PaPrBaG (Deneke et al., 2017) also aims to predict the pathogenic potential of microbial strains by means of training on a large number of established pathogenic species in comparison with non-pathogenic bacteria and their sequence features. PaPrBaG is a RF-based method for the assessment of the pathogenic potential of a set of reads belonging to a single genome. It helps in the prediction of novel, unknown bacterial pathogens. PaPrBaG provides prediction in contrast with other approaches that discard many sequencing reads based on the low similarity to known reference genomes.

3.4. Antimicrobial Resistance Gene Detection

In this step, computational analysis is used to explore whether the previously annotated bacterial genes infer antimicrobial resistance, i.e., the ability of microorganisms to grow despite exposure to antimicrobial substances (Quainoo et al., 2017). However, again, the same is true as for virulence gene prediction—this step does not necessarily have to follow the genome annotation step, e.g., it can be also conducted right after assembly. Frequently used tools for this purpose include a Web-based tool ResFinder (Zankari et al., 2012) and RGI/CARD (Jia et al., 2017). Both perform homology-based resistome prediction: ResFinder (Zankari et al., 2012) uses BLAST, whereas RGI/CARD (Jia et al., 2017) makes use of a manually curated resource containing antimicrobial resistance genes, proteins and mutated sequences—CARD (Jia et al., 2017). Resently, NCBI has developed AMRFinder (Feldgarden et al., 2019) which utilizes the NCBI's curated AMR gene database - Bacterial Antimicrobial Resistance Reference Gene Database-, currently including 4,579 antimicrobial resistance gene proteins and over 560 hidden Markov models (HMMs).

ML approaches for the same task include DeepARG (Arango-Argoty et al., 2018), a deep learning approach using neural networks and previously curated databases, such as CARD (Jia et al., 2017), for predicting antibiotic resistance genes and annotating them to 30 known antibiotic resistance categories, creating a manually curated database, DeepARG-DB. PATRIC (Antonopoulos et al., 2017) uses the genomes in its in-house database and their antimicrobial resistance-related metadata, such as susceptibility or resistance to a given antibiotic, to build AdaBoost (adaptive boosting) ML-based classifiers and predict those regions within a bacterial genome that are associated with antimicrobial resistance (Davis et al., 2016). When a genome is submitted to the PATRIC annotation service, these classifiers are used to predict if the organism is susceptible or resistant to an antibiotic. However, PATRIC is limited to identifying only

genes encoding resistance to certain antibiotics (beta lactam, carbapenem, and methicillin) and in certain bacterial species. In this context, ML has also been applied to identify genomic features possibly related to minimum inhibitory concentration (MIC) of an antibiotic, i.e., its lowest concentration preventing visible growth of bacterium *in vitro*, e.g., for Nontyphoidal *Salmonella* (Nguyen et al., 2019).

4. MACHINE LEARNING FOR MICROBIAL COMPARATIVE GENOMICS

After characterization of an individual genome is accomplished, the next step is to perform comparative genomics and detect relatedness between strains, identify potentially clonal strains and pinpoint the putative source of the outbreak (Brown et al., 2019). Bacterial species should be determined before performing comparative genomic analyses, since most algorithms will perform better when closely related bacterial strains can be used. Comparative genomics methods can be largely divided into three groups: (i) reference/non-reference-based SNP-based methods, (ii) pangenome-based and (iii) core genome/whole-genome multilocus sequence typing (MLST).

4.1. Reference-Based SNP Methods

Standard strategies to identify genetic variation, which occurs in a strain, usually focus on single nucleotide polymorphisms (SNPs). Raw reads are mapped to a perform better when closely related, high-quality reference genome, identifying SNPs as variations in relation to that reference genome. CSI Phylogeny (Kaas et al., 2014), Lyve-SET (Katz et al., 2017), CFSAN SNP Pipeline (Davis et al., 2015), SPANDx (Sarovich and Price, 2014), and SNVPhyl (Petkau et al., 2017) include such pipelines. In addition, there are also tools such as Harvest/Parsnp (Treangen et al., 2014) that, instead of trying to performing whole-genome alignment, focus on constructing a core-genome alignment, i.e., identifying a set of orthologous sequence conserved in all aligned genomes. However, reference-based SNP methods are generally recommended only if a high-quality reference genome exists (Brown et al., 2019), when higher resolution is required than can be achieved using cgMLST/wgMLST, or when a cgMLST/wgMLST scheme is not available (Katz et al., 2017).

4.2. Reference-Free SNP Analysis

Reference-Free SNP Analysis does not require alignment to a reference genome to identify SNPs. Such examples include kSNP (Gardner et al., 2015), a k-mer-based approach where the user provides the length of the flanking sequence including the SNP, i.e., the SNP is at the central base of the k-mer, and the flanking (k-1)/2 bases on both sides of the SNP define the locus. First, kSNP counts all k-mer oligos for each input genome. This is followed by several filtering steps: (i) the k-mer list is then condensed so that counts reflect both occurrences on the forward and reverse strands; (ii) for raw reads, kSNP discards k-mers that occur only once, as such singletons are likely to be sequencing errors; (iii) for each genome, kSNP discards k-mers that have more than one central base variant for a given locus. Finally, kSNP merges and sorts all k-mers across all user provided genomes and looks for SNP loci in the merged list. Then it compares the SNP loci for each genome with the merged list to identify the SNPs in each genome, reporting the locus and the central base, i.e., the SNP, for every genome containing that locus (Gardner et al., 2015).

4.3. Pangenome-Based Analysis

Pangenome-based analysis classifies genes as the so called core genes, found in all bacterial strains under comparison, and into accessory genes that can be found only in several but not all strains (Page et al., 2015). Isolates are then clustered based on their accessory genome (Page et al., 2015). A well-known tool for pangenome-based analysis is Roary (Page et al., 2015). First, it identifies orthologous genes by sequence comparison. This is followed by grouping of these genes into clusters. Finally, the relationships of the clusters are then represented using a graph, constructed based on the order in which their occur in the input data (Page et al., 2015; Brown et al., 2019). Other tools for pangenome-based analysis include PanWeb (Pantoja et al., 2017) and Pan-Seq (Laing et al., 2010).

4.4. Core Genome/Whole-Genome Multi-locus Sequence Typing (MLST)

Core genome/whole-genome multi-locus sequence typing (MLST) are widely used methods for outbreak investigations, enabling standardized outbreak management protocols (Nadon et al., 2017; Brown et al., 2019). Conventional MLST usually uses only seven genes/loci to derive sequence types (STs), and is not always able to distinguish between outbreaks resulting from closely related bacterial variants (Pearce et al., 2018). Core genome MLST (cgMLST) schemes extend the conventional MLST, including genes/loci present in 95% to 99% of isolates, hence offering increased resolution to detect isolate-specific genotypes, as well as novel transmission events (Nadon et al., 2017; Brown et al., 2019). If two strains display identical cgMLST profiles, these are being grouped into one cluster type (CT), which can be shared using dedicated databases (Quainoo et al., 2017). CgMLST is implemented within the Ridom SeqSphere+ commercial software suite (JÃijnemann et al., 2013). However, it is also being utilized by EnteroBase (Alikhan et al., 2018), Bacterial Isolate Genome Sequence Database (BIGSdb) (Jolley and Maiden, 2010) and chewBBACA (Silva et al., 2018). On the other hand, whole-genome MLST (wgMLST) further extends cgMLST, as it also considers the accessory genes to detect lineage-specific loci. This method is part of the BioNumerics (Applied Maths) software suite since version 7.5 (http://www.applied-maths.com/) and is also implemented within EnteroBase (Alikhan et al., 2018). For outbreak investigations, cgMLST is more suited, as it uses species-specific nomenclature; however, wgMLST might offer higher resolution to discriminate outbreak strains that form closely related clusters (Nadon et al., 2017; Brown et al., 2019). Of note, however, both methods strongly depend on the availability of high-resolution isolate typing schemes (Pearce et al., 2018), which may not be available for lesser-studied foodborne pathogens, due to the lack of publicly available WGS data (Carroll et al., 2019).

To the best of our knowledge, ML-based tools do not seem to have gained a lot of attention in comparative genomics. The Bayesian Analysis of Population Structure (BAPS)/hierBAPS

(Cheng et al., 2011, 2013) tool seems to be the only ML-based tool for comparative genomics. BAPS/hierBAPS was created by first collecting large data sets of multi-locus DNA sequence types (STs), as well as the respective metadata (e.g., host organism, serotype) from several MLST databases PubMLST (http://www.pubmlst.org). This data was then utilized to divide the available pathogens into subsets of different evolutionary lineages or geographically related sub-populations, as determined based on molecular [dis]similarities within the database. Then a user-submitted set of bacterial isolates can be classified to one of these groups, using a Bayesian model-based ML algorithm. In addition, recently, several other studies have combined comparative genomics with ML approaches for the classification of outbreak strains (Diaz et al., 2017) or source tracking during outbreaks (Buultjens et al., 2017; Zhang et al., 2019). Diaz et al. (2017) identified six distinct subtypes of genomes, as well as their respective SNPs/loci, and trained RF to separate input genomes into the respective subtypes. Buultjens et al. (2017) used core genome variation and classification based on principal components to identify genomic signatures specific to source of interest, which were further used to predict the origin of input isolates (Buultjens et al., 2017). Zhang et al. (2019) used a set of genetic features extracted from *Salmonella* Typhimurium genomes, inlcuding core genome SNPs, insertion/deletions and accessory genes to train a RF classifier in discriminating isolates from swine, bovine, poultry or wild bird sources. Wheeler et al. (2018) investigated genomic signatures related to host adaptation in *Salmonella enterica*. First, hidden Markov models were used to identify patterns of sequence variation and their potential functional consequences. Thereafter, RF was utilized to identify genes that displayed differences between lineages with different phenotypes (Wheeler et al., 2018). Sharma et al. (2014) used MLST to differentiate isolates and categorize an unknown isolate as either representing a true infection or a likely contaminant. In particular, the seven genotypes derived from MLST were used to train three different ML algorithms (SVM; Classification And Regression Tree Analysis - CART; and a Naive Nearest-Neighbor Classifier) to segregate isolates of known class (i.e., pathogen or likely contaminant) on the basis of their alleles, which were then used to classify an unknown isolate by its MLST allele profile.

5. MACHINE LEARNING FOR THE INFERENCE OF MICROBIAL PHYLOGENOMICS

Finally, comparison tools can be used for the inference of microbial phylogenomics of pathogenic isolates and generate detailed networks reflecting the transmission events of outbreak strains between different patients (Quainoo et al., 2017). In particular, phylogenomics can reveal whether two isolates are nearly identical or only distantly related and which might represent the initial outbreak source strain (Quainoo et al., 2017). Maximum likelihood is frequently applied when characterizing pathogens from foodborne outbreaks. RAxML (Randomized Axelerated Maximum Likelihood) (Stamatakis et al., 2005) and FastTree (Price et al., 2009) are two maximum likelihood based phylogenomics estimators, which work by first constructing an initial tree, which is then further refined in several optimization steps and tree rearrangements to increase the likelihood that the respective tree reflects the evolutionary relationships of the input sequences. These software packages are often included in the genome comparison pipelines mentioned in the previous chapter such as CSI Phylogeny (Kaas et al., 2014) and Lyve-SET (Katz et al., 2017) for streamlined production of actionable results. Alternatively, distance matrix-based methods such as neighbor joining (Saitou and Nei, 1987) (e.g., part of the PHYLIP Shimada and Nishida, 2017 package) as well as Bayesian analysis-based methods (e.g., BEAST Drummond and Rambaut, 2007) have been proposed to study microbial phylogenomics.

Most recently, Suvorov et al. (2019) has proposed an approach that uses convolutional neural networks (CNNs) for phylogenetic inference. In particular, CNNs are being trained to extract phylogenetic signal from a multiple sequence alignment, which is then used to reconstruct and discriminate alternative tree topologies. Of note, however, this study used an alignment of only four sequences.

6. CONCLUSIONS

Over the last years, several ML-based tools have been developed for different steps of bacterial WGS analysis. However, some areas of bacterial bioinformatics (i.e., genome assembly and strain identification) have seen more development than others (i.e., phylogeny estimation). Overall, AI and its sub-discipline ML could lead to actionable knowledge in diverse ranges of sectors, where multiple complex challenges need to be addressed, including the outbreak investigations of foodborne pathogens and antimicrobial resistance (Gibbs, 2014; Quainoo et al., 2017; Ching et al., 2018), considering that WGS may replace conventional analysis methods already in the near future (Quainoo et al., 2017). In this scenario, the success of outbreak investigations will largely depend on how fast and accurate WGS data can be produced and analyzed (Quainoo et al., 2017). ML-based algorithms could further speed-up such investigations, especially as the number of complete microbial genomes in NCBI RefSeq (http://www.ncbi.nlm.nih.gov/genome) is rapidly growing (Tatusova et al., 2015), providing a valuable resource for training ML classifiers. However, even if substantially improving the accuracy and speed of WGS algorithms, a number of limitations still need to be overcome in order to fully utilize the power of ML for outbreak screenings. WGS analysis tools often rely on sequence similarity and hence strongly depend on reference databases (Deneke et al., 2017; Zhang et al., 2017). Moreover, such methods are rather time-consuming and computationally demanding, thus representing a bottleneck for efficient sequence data analysis (Sharma et al., 2015). ML algorithms could potentially increase the accuracy and speed of clinically and epidemiologically relevant predictions (Farrell et al., 2018). However, to yield accurate predictions, besides the choice of the most appropriate algorithm and a set of well-defined inputs and outputs of

interest, ML-based strategies generally require large amounts of high-quality training data (Baker et al., 2018). This presents a limitation, as currently microbial genome databases are known to be biased toward cultivable pathogenic bacteria. The current lack of large and comprehensive databases can be considered as the key bottleneck for the application of ML methods (Farrell et al., 2018). Hence, future improvements can be expected to come from better data curation and collection, in addition to development of new and improved classification algorithms (Farrell et al., 2018). Therefore, WGS data collection must be done in parallel with comprehensive and standartized metadata collection such as phenotypic profiling using traditional microbiology methods for isolate characterization (e.g., phenotypic profiling of antimicrobial resistance) (Maurer et al., 2017).

Currently, sequencing of bacterial genomes is mostly performed on Illumina instruments, producing relatively short reads with limited resolution of low-complexity regions (Quainoo et al., 2017). Alternatively, ultra-long read technologies such as ONT (https://nanoporetech.com/) and PacBio SMRT (https://www.pacb.com/smrt-science/smrt-sequencing/) are increasingly being used to obtain complete microbial genomes. However, both technologies are still three and almost seven times more expensive in comparison to Illumina short-read sequencing (Brown et al., 2017; Sekse et al., 2017; Nicola De Maio, 2019). Moreover, both technologies still display rather high error rates (Mahmoud et al., 2017), which makes them more suitable for gap closure in draft genomes using hybrid methods (Quainoo et al., 2017). Hence, error-profile-aware ML-algorithms implementing hybrid strategies that make use of more accurate short reads in conjunction with ultra-long reads may need to be considered for future applications.

The selection of a harmonized bioinformatics strategy or pipeline that would perform consistently across outbreak investigation situations around the world, reaching consensus on desired standards represents another challenge for the routine implementation of WGS analysis (Quainoo et al., 2017). Especially, considering that the numbers of commercial analysis software platforms, as well as open-source, application-specific analysis tools are increasing, a rigorous assessment and benchmarking of their quality is urgently needed (Quainoo et al., 2017). This would also be a prerequisite for a systematic comparison between ML-based vs. conventional methods. Nevertheless, in order to perform such comparisons on a global scale, WGS data storage and sharing would be of utmost importance. Although technically feasible, this will require us to solve several issues of ownership and data privacy, making sure that these are being adequately protected (Quainoo et al., 2017).

AUTHOR CONTRIBUTIONS

BV wrote the manuscript. IM, LG-I, and JK participated in revising and editing the manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

This research was funded by the ERDF and state budget co-financed project No. 1.1.1.1/16/A/258 "Development and the application of innovative instrumental analytical methods for the combined determination of a wide range of chemical and biological contaminants in support of the bio-economy in the priority sectors of economy".

REFERENCES

Afiahayati, Sato, K., and Sakakibara, Y. (2015). Metavelvet-sl: an extension of the velvet assembler to a *de novo* metagenomic assembler utilizing supervised learning. *DNA Res.* 22, 69–77. doi: 10.1093/dnares/dsu041

Afify, H. M., and Al-Masni, M. A. (2018). Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches. *Inform. Med. Unlocked* 13, 151–157. doi: 10.1016/j.imu.2018.05.004

Alikhan, N.-F., Zhou, Z., Sergeant, M. J., and Achtman, M. (2018). A genomic overview of the population structure of salmonella. *PLoS Genet.* 14:e1007261. doi: 10.1371/journal.pgen.1007261

Alkema, W., Boekhorst, J., Wels, M., and van Hijum, S. A. F. T. (2016). Microbial bioinformatics for food safety and production. *Brief. Bioinformat.* 17, 283–292. doi: 10.1093/bib/bbv034

Andersen, S. C., and Hoorfar, J. (2018). Surveillance of foodborne pathogens: Towards diagnostic metagenomics of fecal samples. *Genes* 9:E14. doi: 10.3390/genes9010014

Antonopoulos, D. A., Assaf, R., Aziz, R. K., Brettin, T., Bun, C., Conrad, N., et al. (2017). Patric as a unique resource for studying antimicrobial resistance. *Brief. Bioinform*. doi: 10.1093/bib/bbx083. [Epub ahead of print].

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. doi: 10.1186/s40168-018-0401-z

Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4fun: predicting functional profiles from metagenomic 16s rrna data. *Bioinformatics* 31, 2882–2884. doi: 10.1093/bioinformatics/btv287

Baker, R. E., Peña, J.-M., Jayamohan, J., and Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.* 14:20170660. doi: 10.1098/rsbl.2017.0660

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bergholz, T. M., Moreno Switt, A. I., and Wiedmann, M. (2014). Omics approaches in food safety: fulfilling the promise? *Trends Microbiol.* 22, 275–281. doi: 10.1016/j.tim.2014.01.006

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* 17, 1519–1533. doi: 10.1089/cmb.2009.0238

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Pittsburgh, PA: ACM), 144–152.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140.

Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C., and Franklin, R. B. (2017). MinION^{TM} nanopore sequencing of environmental metagenomes: a synthetic approach. *GigaScience* 6, 1–10. doi: 10.1093/gigascience/gix007

Brown, E., Dessai, U., McGarry, S., and Gerner-Smidt, P. (2019). Use of whole-genome sequencing for food safety and public health in the united states. *Foodborne Pathogens Dis.* 16, 441–450. doi: 10.1089/fpd.2019.2662

Buultjens, A. H., Chua, K. Y. L., Baines, S. L., Kwong, J., Gao, W., Cutcher, Z., et al. (2017). A supervised statistical learning approach for accurate Legionella

pneumophila source attribution during outbreaks. *Appl. Environ. Microbiol.* 83: e01482-17. doi: 10.1128/AEM.01482-17

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carroll, L. M., Wiedmann, M., Mukherjee, M., Nicholas, D. C., Mingle, L. A., Dumas, N. B., et al. (2019). Characterization of emetic and diarrheal bacillus cereus strains from a 2016 foodborne outbreak using whole-genome sequencing: addressing the microbiological, epidemiological, and bioinformatic challenges. *Front. Microbiol.* 10:144. doi: 10.3389/fmicb.2019.00144

CDC (2013). *Antibiotic Resistance Threats in the United States, 2013*. Technical report, CDC. Atlanta, GA.

Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A., and Sharma, V. K. (2015). 16s classifier: a tool for fast and accurate taxonomic classification of 16s rrna hypervariable regions in metagenomic datasets. *PLoS ONE* 10:e0116106. doi: 10.1371/journal.pone.0116106

Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). Vfdb 2016: hierarchical and refined dataset for big data analysis–10 years on. *Nucleic Acids Res.* 44, D694–D697. doi: 10.1093/nar/gkv1239

Cheng, L. (2015). *A Machine Learning Approach to DNA Shotgun Sequence Assembly*. PhD thesis, University of the Witwatersrand. Johannesburg, South Africa.

Cheng, L., Connor, T. R., Aanensen, D. M., Spratt, B. G., and Corander, J. (2011). Bayesian semi-supervised classification of bacterial samples using MLST databases. *BMC Bioinform.* 12:302. doi: 10.1186/1471-2105-12-302

Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M., and Corander, J. (2013). Hierarchical and spatially explicit clustering of dna sequences with baps software. *Mol. Biol. Evol.* 30, 1224–1228. doi: 10.1093/molbev/mst028

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15: 20170387. doi: 10.1098/rsif.2017.0387

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., et al. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* 68, 2386–2392. doi: 10.1099/ijsem.0.002809

Clingenpeel, S., Clum, A., Schwientek, P., Rinke, C., and Woyke, T. (2015). Reconstructing each cell's genome within complex microbial communities' dream or reality? *Front. Microbiol.* 5:771. doi: 10.3389/fmicb.2014.00771

Cosentino, S., Voldby Larsen, M., Møller Aarestrup, F., and Lund, O. (2013). Pathogenfinder–distinguishing friend from foe using bacterial whole genome sequence data. *PLoS ONE* 8:e77302. doi: 10.1371/journal.pone.0077302

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., et al. (2016). Antimicrobial resistance prediction in patric and rast. *Sci. Rep.* 6:27930. doi: 10.1038/srep27930

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., et al. (2015). Cfsan snp pipeline: an automated method for constructing snp matrices from next-generation sequence data. *PeerJ Comput. Sci.* 1:e20. doi: 10.7717/peerj-cs.20

Deneke, C., Rentzsch, R., and Renard, B. Y. (2017). Paprbag: a machine learning approach for the detection of novel pathogens from NGS data. *Sci. Rep.* 7:39194. doi: 10.1038/srep39194

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A Probabilistic Theory of Pattern Recognition*, Vol 31. Springer Science & Business Media.

Diaz, M. H., Desai, H. P., Morrison, S. S., Benitez, A. J., Wolff, B. J., Caravas, J., et al. (2017). Comprehensive bioinformatics analysis of mycoplasma pneumoniae genomes to investigate underlying population structure and type-specific determinants. *PLoS ONE* 12:e0174701. doi: 10.1371/journal.pone.0174701

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., and Nattkemper, T. W. (2009). Tacoa: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinform.* 10:56. doi: 10.1186/1471-2105-10-56

Drummond, A. J., and Rambaut, A. (2007). Beast: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Dutilh, B. E., Backus, L., Edwards, R. A., Wels, M., Bayjanov, J. R., and van Hijum, S. A. F. T. (2013). Explaining microbial phenotypes on a genomic scale: Gwas for microbes. *Brief. Funct. Genom.* 12, 366–380. doi: 10.1093/bfgp/elt008

EFSA (2015). The european union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in (2014). *EFSA J.* 13:191. doi: 10.2903/j.efsa.2015.4329

El Allali, A., and Rose, J. R. (2013). Mgc: a metagenomic gene caller. *BMC Bioinform.* 14 (Suppl. 9):S6. doi: 10.1186/1471-2105-14-S9-S6

Farrell, F., Soyer, O. S., and Quince, C. (2018). Machine learning based prediction of functional capabilities in metagenomically assembled microbial genomes. *bioRxiv*. doi: 10.1101/307157

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., et al. (2019). Using the ncbi amrfinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of narms isolates. *bioRxiv*. doi: 10.1101/550707

Gao, X., Lin, H., Revanna, K., and Dong, Q. (2017). A bayesian taxonomic classification method for 16s rRNA gene sequences with improved species-level accuracy. *BMC Bioinform.* 18:247. doi: 10.1186/s12859-017-1670-4

Gardner, S. N., Slezak, T., and Hall, B. G. (2015). ksnp3.0: Snp detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31, 2877–2878. doi: 10.1093/bioinformatics/btv271

Garg, A., and Gupta, D. (2008). Virulentpred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinform.* 9:62. doi: 10.1186/1471-2105-9-62

Gibbs, E. P. J. (2014). The evolution of one health: a decade of progress and challenges for the future. *Veter. Record* 174, 85–91. doi: 10.1136/vr.g143

Gregor, I., Dröge, J., Schirmer, M., Quince, C., and McHardy, A. C. (2016). Phylopythias+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* 4:e1603. doi: 10.7717/peerj.1603

Han, N., Qiang, Y., and Zhang, W. (2016). Anitools web: a web tool for fast genome comparison within multiple bacterial strains. *Database* 2016: baw084. doi: 10.1093/database/baw084

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. *J. R. Statist. Soc. Ser. C* 28, 100–108.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C. A., Frimodt-MÃ¿ller, N., et al. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol.* 52, 139–146. doi: 10.1128/JCM.02452-13

Hendriksen, R. S., Pedersen, S. K., Leekitcharoenphon, P., Malorny, B., Borowiak, M., Battisti, A., et al. (2018). Final report of engage-establishing next generation sequencing ability for genomic analysis in europe. *EFSA Supp. Public.* 15:1431E. doi: 10.2903/sp.efsa.2018.EN-1431

Her, H.-L., and Wu, Y.-W. (2018). A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. *Bioinformatics* 34, i89–i95. doi: 10.1093/bioinformatics/bty276

Hoff, K. J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 37, W101–W105. doi: 10.1093/nar/gkp327

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24:417.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* 11:119. doi: 10.1186/1471-2105-11-119

Iraola, G., Vazquez, G., Spangenberg, L., and Naya, H. (2012). Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans. *PLoS ONE* 7:e42144. doi: 10.1371/journal.pone.0042144

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., et al. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* 52, 1501–1510. doi: 10.1128/JCM.03617-13

Jolley, K. A., and Maiden, M. C. (2010). Bigsdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform.* 11:595. doi: 10.1186/1471-2105-11-595

Jünemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., et al. (2013). Updating benchtop sequencing performance comparison. *Nat. Biotechn.* 31, 294–296. doi: 10.1038/nbt.2522

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M., and Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS ONE* 9:e104984. doi: 10.1371/journal.pone.0104984

Kanamori, H., Parobek, C. M., Weber, D. J., van Duin, D., Rutala, W. A., Cairns, B. A., et al. (2015). Next-generation sequencing and comparative analysis of sequential outbreaks caused by multidrug-resistant acinetobacter baumannii at a large academic burn center. *Antimicrob. Agents. Chemother.* 60, 1249–1257. doi: 10.1128/AAC.02014-15

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., et al. (2017). A comparative analysis of the lyve-set phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Front. Microbiol.* 8:375. doi: 10.3389/fmicb.2017.00375

Kolbe, D. L., and Eddy, S. R. (2011). Fast filtering for rna homology search. *Bioinformatics* 27, 3102–3109. doi: 10.1093/bioinformatics/btr545

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive, javax.xml.bind.jaxbelement@19c8c323, -mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., and Steinbrecher, M. (2016). *Computational Intelligence: A Methodological Introduction*. Heidelberg: Springer.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., and Ussery, D. W. (2007). Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Lai, K., Twine, N., O'Brien, A., Guo, Y., and Bauer, D. (2016). "Artificial intelligence and machine learning in bioinformatics," in *Encyclopedia of Bioinformatics and Computational Biology*, ed M. Gribskov (Elsevier). doi: 10.1016/B978-0-12-809633-8.20325-7

Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., et al. (2010). Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinform.* 11:461. doi: 10.1186/1471-2105-11-461

Laslett, D., and Canback, B. (2004). Aragorn, a program to detect trna genes and tmrna genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi: 10.1093/nar/gkh152

Lee, I., Ouk Kim, Y., Park, S.-C., and Chun, J. (2016). Orthoani: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* 66, 1100–1103. doi: 10.1099/ijsem.0.000760

Li, H. (2016). Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 32, 2103–2110. doi: 10.1093/bioinformatics/btw152

Li, L.-G., Yin, X., and Zhang, T. (2018). Tracking antibiotic resistance gene pollution from different sources using machine-learning classification. *Microbiome* 6:93. doi: 10.1186/s40168-018-0480-x

Liu, Y., Guo, J., Hu, G., and Zhu, H. (2013). Gene prediction in metagenomic fragments based on the svm algorithm. *BMC Bioinform.* 14 (Suppl. 5):S12. doi: 10.1186/1471-2105-14-S5-S12

Llarena, A.-K., Ribeiro-Gonçalves, B. F., Nuno Silva, D., Halkilahti, J., Machado, M. P., Da Silva, M. S., et al. (2018). Innuendo: A cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. *EFSA Supp. Public.* 15:1498E. doi: 10.2903/sp.efsa.2018.EN-1498

Lupolova, N., Dallman, T. J., Holden, N. J., and Gally, D. L. (2017). Patchy promiscuity: machine learning applied to predict the host specificity of salmonella enterica and *Escherichia coli*. *Microb. Genom.* 3:e000135. doi: 10.1099/mgen.0.000135

Mahmoud, M., Zywicki, M., Twardowski, T., and Karlowski, W. M. (2017). Efficiency of pacbio long read correction by 2nd generation illumina sequencing. *Genomics* 111, 43–49. doi: 10.1016/j.ygeno.2017.12.011

Maurer, F. P., Christner, M., Hentschke, M., and Rohde, H. (2017). Advances in rapid identification and susceptibility testing of bacteria in the clinical microbiology laboratory: implications for patient care and antimicrobial stewardship programs. *Infect. Dis. Rep.* 9:6839. doi: 10.4081/idr.2017.6839

McGinnis, S., and Madden, T. L. (2004). Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32, W20–W25. doi: 10.1093/nar/gkh435

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length dna fragments. *Nat. Methods* 4, 63–72. doi: 10.1038/nmeth976

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9:386. doi: 10.1186/1471-2105-9-386

Moran-Gilad, J. (2017). Whole genome sequencing (wgs) for food-borne pathogen surveillance and control - taking the pulse. *Euro Surveill.* 22:30547. doi: 10.2807/156

Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., et al. (2017). Pulsenet international: vision for the implementation of whole genome sequencing (wgs) for global food-borne disease surveillance. *Euro Surveill.* 22: 30544. doi: 10.2807/1560-7917.ES.2017.22.23.30544

Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., et al. (2019). Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* 57: e01260-18. doi: 10.1128/JCM.01260-18

Nicola De Maio, Liam P. Shaw, A. H. S. G. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *bioRxiv* doi: 10.1101/530824

Noguchi, H., Park, J., and Takagi, T. (2006). Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396. doi: 10.1093/dnares/dsn027

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Palmer, L. E., Dejori, M., Bolanos, R., and Fasulo, D. (2010). Improving *de novo* sequence assembly using machine learning and comparative genomics for overlap correction. *BMC Bioinform.* 11:33. doi: 10.1186/1471-2105-11-33

Pantoja, Y., Pinheiro, K., Veras, A., AraÃžjo, F., Lopes de Sousa, A., GuimarÃčes, L. C., et al. (2017). Panweb: A web interface for pan-genomic analysis. *PLoS ONE* 12:e0178154. doi: 10.1371/journal.pone.0178154

Pearce, M. E., Alikhan, N.-F., Dallman, T. J., Zhou, Z., Grant, K., and Maiden, M. C. J. (2018). Comparative analysis of core genome mlst and snp typing within a european salmonella serovar enteritidis outbreak. *Int. J. Food Microbiol.* 274, 1–11. doi: 10.1016/j.ijfoodmicro.2018.02.023

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). Idba-ud: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Petkau, A., Mabon, P., Sieffert, C., Knox, N. C., Cabral, J., Iskander, M., et al. (2017). Snvphyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb. Genom.* 3:e000116. doi: 10.1099/mgen.0.000116

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Quainoo, S., Coolen, J. P., van Hijum, S. A., Huynen, M. A., Melchers, W. J., van Schaik, W., et al. (2017). Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clin. Microbiol. Rev.* 30, 1015–1063. doi: 10.1128/CMR.00016-17

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of salmonella. *Genome Biol.* 16:114. doi: 10.1186/s13059-015-0677-2

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. (2016). Jspeciesws: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32, 929–931. doi: 10.1093/bioinformatics/btv681

Roosaare, M., Vaher, M., Kaplinski, L., Möls, M., Andreson, R., Lepamets, M., et al. (2017). Strainseeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ* 5:e3353. doi: 10.7717/peerj.3353

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using n-mer frequency profiles. *Adv. Bioinform.* 2008:205969. doi: 10.1155/2008/205969

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Sarovich, D. S., and Price, E. P. (2014). Spandx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. *BMC Res. Notes* 7:618. doi: 10.1186/1756-0500-7-618

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Sedlar, K., Kupkova, K., and Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotech. J.* 15, 48–55. doi: 10.1016/j.csbj.2016.11.005

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Sekse, C., Holst-Jensen, A., Dobrindt, U., Johannessen, G. S., Li, W., Spilsberg, B., et al. (2017). High throughput sequencing for detection of foodborne pathogens. *Front. Microbiol.* 8:2029. doi: 10.3389/fmicb.2017.02029

Sharma, A. K., Gupta, A., Kumar, S., Dhakan, D. B., and Sharma, V. K. (2015). Woods: a fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics* 106, 1–6. doi: 10.1016/j.ygeno.2015.04.001

Sharma, P., Satorius, A. E., Raff, M. R., Rivera, A., Newton, D. W., and Younger, J. G. (2014). Multilocus sequence typing for interpreting blood isolates of staphylococcus epidermidis. *Int. Perspect. Infect. Dis.* 2014:787458. doi: 10.1155/2014/787458

Shimada, M. K., and Nishida, T. (2017). A modification of the phylip program: a solution for the redundant cluster problem, and an implementation of an automatic bootstrapping on trees inferred from original data. *Mol. Phylogenet. Evol.* 109, 409–414. doi: 10.1016/j.ympev.2017.02.012

Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., et al. (2018). chewbbaca: a complete suite for gene-by-gene schema creation and strain identification. *Microb. Genom.* 4. doi: 10.1099/mgen.0.000166

Souvorov, A., Agarwala, R., and Lipman, D. J. (2018). Skesa: strategic k-mer extension for scrupulous assemblies. *Genome Biol.* 19:153. doi: 10.1186/s13059-018-1540-z

Stamatakis, A., Ludwig, T., and Meier, H. (2005). Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21, 456–463. doi: 10.1093/bioinformatics/bti191

Suvorov, A., Hochuli, J., and Schrider, D. (2019). Accurate inference of tree topologies from multiple sequence alignments using deep learning. *bioRxiv*. doi: 10.1101/559054

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., et al. (2015). Update on refseq microbial genomes resources. *Nucleic Acids Res.* 43, D599–D605. doi: 10.1093/nar/gku1062

Tebani, A., Afonso, C., Marret, S., and Bekri, S. (2016). Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int. J. Mol. Sci.* 17: E1555. doi: 10.3390/ijms17091555.

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/PREACCEPT-2573980311437212

Vangay, P., Steingrimsson, J., Wiedmann, M., and Stasiewicz, M. J. (2014). Classification of listeria monocytogenes persistence in retail delicatessen environments using expert elicitation and machine learning. *Risk Anal.* 34, 1830–1845. doi: 10.1111/risa.12218

Wheeler, N. E., Gardner, P. P., and Barquist, L. (2018). Machine learning identifies signatures of host adaptation in the bacterial pathogen salmonella enterica. *PLoS Genet.* 14:e1007333. doi: 10.1371/journal.pgen.1007333

WHO (2014). *Antimicrobial Resistance: Global Report on Surveillance.* Technical report, WHO.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Yuan, C., Lei, J., Cole, J., and Sun, Y. (2015). Reconstructing 16s rrna genes in metagenomic data. *Bioinformatics* 31, i35–i43. doi: 10.1093/bioinformatics/btv231

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi: 10.1093/jac/dks261

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhang, S., Li, S., Gu, W., den Bakker, H., Boxrud, D., Taylor, A., et al. (2019). Zoonotic source attribution of salmonella enterica serotype typhimurium using genomic surveillance data, united states. *Emerg. Infect. Dis.* 25, 82–91. doi: 10.3201/eid2501.180835

Zhang, S.-W., Jin, X.-Y., and Zhang, T. (2017). Gene prediction in metagenomic fragments with deep learning. *BioMed Res. Int.* 2017:4740354. doi: 10.1155/2017/4740354

Zhu, X., Leung, H. C. M., Chin, F. Y. L., Yiu, S. M., Quan, G., Liu, B., et al. (2014). Perga: a paired-end read guided *de novo* assembler for extending contigs using svm and look ahead approach. *PLoS ONE* 9:e114253. doi: 10.1371/journal.pone.0114253

Conflict of Interest Statement: BV is the CEO of net-OMICS, a bioinformatics company.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.